

**AMENDMENTS TO THE CLAIMS:**

The listing of claims will replace all prior versions, and listings of claims in the application:

**LISTING OF CLAIMS:**

1. (Cancelled)

2. (Cancelled)

3. (Currently amended) A method performed by a computer for clustering a plurality of documents in a structure comprised of a plurality of clusters hierarchically organized, wherein each document includes a plurality of words and is represented as a set of (document, word) pairs, the method comprising:

\_\_\_\_\_ accessing the document collection;

\_\_\_\_\_ performing a clustering process that creates a hierarchy of clusters that reflects a segregation of the documents in the collection based on the words included in the documents, wherein a document in the collection is assigned to a first cluster in the hierarchy based on a first segment of the respective document, and the respective document is assigned to a second cluster in the hierarchy based on a second segment of the respective document, wherein the first and second clusters are associated with different paths of the hierarchy The method of claim 2, wherein the step of determining further comprises: comprising the steps:

\_\_\_\_\_ assigning the document collection to a first class;

\_\_\_\_\_ setting a probability parameter to an initial value;

\_\_\_\_\_ determining, for each document in the collection at the value of the

parameter, a probability of an assignment of the document in the collection to a cluster in the hierarchy based on a word included in the document and the first class;

\_\_\_\_\_determining whether the first class has split into two child classes, wherein each child class reflects a cluster descendant from an initial cluster reflected by the first class; and

\_\_\_\_\_increasing the value of the parameter based on the determination whether the first class has split into two child classes.

\_\_\_\_\_storing a representation of the hierarchy of clusters in a memory; and

\_\_\_\_\_making the representation available to an entity in response to a request associated with the document collection.

4. (Original) The method of claim 3, further comprising:

repeating the step of determining, for each document in the collection at the value of the parameter, and the step of increasing the value of the parameter until the first class has split into two child classes.

5. (Original) The method of claim 4, further comprising:

performing the clustering process for each child class until each of the respective child class splits into two new child classes reflecting clusters descendant from the respective child class.

6. (Original) The method of claim 5, further comprising:

repeating the clustering process for each new child class such that a hierarchy of

clusters is created, until a predetermined condition associated with the hierarchy is met.

7. (Original) The method of claim 6, wherein the predetermined condition is one of a maximum number of leaves associated with the hierarchy and depth level of the hierarchy.

8. (Cancelled)

9. (Currently amended) A method performed by a computer for determining topics of a document collection, the method comprising:

\_\_\_\_\_ accessing the document collection, each document including a plurality of words and being represented as a set of (document, word) pairs;

\_\_\_\_\_ performing a clustering process including:

\_\_\_\_\_ creating a tree of nodes that represent topics associated with the document collection based on the words in the document collection, wherein at least one node in the tree includes a word that is shared by another node in the tree, and

\_\_\_\_\_ assigning fragments of one or more documents included in the document collection to multiple nodes in the tree based on the (document, word) pairs comprising: The method of claim 8, wherein the step of assigning comprises:

\_\_\_\_\_ associating a set of documents in the document collection with a first class reflecting all of the nodes in the tree, wherein the set of documents may include all or some of the documents in the collection;

\_\_\_\_\_ defining a second class reflecting any ancestor node of a node in

the first class;

\_\_\_\_\_ determining, for each document in the set, a probability that different words included in a respective document co-occurs with the respective document in any node in the tree based on the first and second classes; and

\_\_\_\_\_ assigning one or more fragments of any document in the set to any node in the tree based on the probability;

\_\_\_\_\_ storing a representation of the tree in a memory; and

\_\_\_\_\_ making the representation available for processing operations associated with the document collection.

10. (Currently amended) A computer program product, comprising a computer usable medium having a computer readable program code embodied therein, the computer readable program code adapted to be executed by a computer to implement a method  
A method performed by a processor for clustering data in a database, the method comprising:

receiving a collection of documents, each document including a plurality of words and being represented as a set of (document, word) pairs;

creating a first ancestor node reflecting a first topic based on words included in the collection of documents;

creating descendant nodes from the first ancestor node, each descendant node reflecting descendant topics based on the first node, until a set of leaf nodes reflecting leaf topics are created,

wherein creating descendant nodes includes:

assigning each document in the collection to a plurality of descendant and leaf nodes; and

providing a set of topics associated with the collection of documents based on the created nodes and assignment of documents,

wherein the descendant and leaf nodes may be created based on one or more words included in more than one document in the collection of documents, and

wherein a probability of observing any pair of co-occurring objects represented as the set of (document,word) pairs, the document of the pairs being (i) and the words being (j), is modeled by defining a variable  $I_{i\alpha}$  which controls assignment of documents to the hierarchy, such that it is dependent on a particular (document,word) pair (i,j) under consideration during a topical clustering process, and a class  $\alpha$  ranges over all nodes in an induced hierarchy in order to assign a document (i object) to any node in the hierarchy, not just leaf nodes, and a class  $v$  is defined as an ancestor of  $\alpha$  in the hierarchy, the constraint on  $v$  ensuring that the nodes are hierarchically organized.

11. (Currently amended) The computer program product The method of claim 10, wherein the step of creating descendant nodes comprises:

selecting a first document in the collection; defining a first class that includes all of the nodes;

defining a second class that may include any ancestor node of any node included in the first class; and

determining, for each document in the collection, a target word of an object pair including a target document and the target word such that the first document equals the

target document in the object pair based on a probability associated with the first and second classes; and

assigning the first document to any ancestor, descendant, and leaf node based on the determining.

12. (Currently amended) A computer program product, comprising a computer usable medium having a computer readable program code embodied therein, the computer readable program code adapted to be executed by a computer to implement a method  
A method performed by a processor for clustering data in a database, the method  
comprising:

receiving a collection of documents, each document including a plurality of words and being represented as a set of (document, word) pairs, the document of the pairs being (i) and the words of the pairs being (j);

creating a hierarchy of nodes based on the words in the collection of documents, each node reflecting a topic associated with the documents, wherein the hierarchy of nodes includes ancestor nodes, descendant nodes, and leaf nodes;

assigning each document in the collection to a plurality of nodes in the hierarchy, wherein each document has the ability to be assigned to any of the ancestor, descendant, and leaf nodes; and

providing a set of topic clusters associated with the collection of documents based on the created nodes and assignment of documents,

wherein the hierarchy may include a plurality of nodes that are each created based on a same set of words included in the collection of documents,

wherein in the foregoing steps, the (j) objects, for a given  $\alpha$ , are not collected in a product, rather a probability is determined such that the product is taken only after mixing over all classes  $\alpha$ , thus, different j objects are generated from different vertical paths of an induced hierarchy, that is, the paths in the hierarchy associated with non-null values of  $l_{i\alpha}$ , and all the instances of the hidden variable  $l_{i\alpha}$  obtain real values after a re-estimation using a modified EM process, and wherein as  $\alpha$  is able to be any node in the hierarchy, the (i) objects are able to be assigned to different levels of the hierarchy, accordingly, implementation of the model results in a pure soft hierarchical clustering of both (i) and (j) objects by eliminating any hard assignments of these objects.

13. (Currently amended) A method performed by a computer for clustering data stored on a computer-readable medium, the method comprising:

receiving a collection of data objects, represented as a set of (first data object, second data object) pairs, the document of the pairs being (i) and the words of the pairs being (j);

for each first data object:

assigning the first data object to a first node in a hierarchy of nodes based on the second data objects included in the first data object, wherein the first node is any node included in the hierarchy and wherein two or more nodes in the hierarchy share the same second object;

creating a final hierarchy of nodes by successively splitting ancestor nodes arranged in clusters based on the assignment of the first data objects;

storing a representation of the final hierarchy in a memory; and

making the representation of the final hierarchy available to an entity in response to a request associated with the collection of first data objects,

wherein in the foregoing steps, the (j) objects, for a given  $\alpha$ , are not collected in a product, rather a probability is determined such that the product is taken only after mixing over all classes  $\alpha$ , thus, different j objects are generated from different vertical paths of an induced hierarchy, that is, the paths in the hierarchy associated with non-null values of  $l_{i\alpha}$ , and all the instances of the hidden variable  $l_{i\alpha}$  obtain real values after a re-estimation using a modified EM process, and wherein as  $\alpha$  is able to be any node in the hierarchy, the (i) objects are able to be assigned to different levels of the hierarchy, accordingly, implementation of the model results in a pure soft hierarchical clustering of both (i) and (j) objects by eliminating any hard assignments of these objects.

14. (Currently amended) A method performed by a processor for clustering data in a database, the method comprising:

receiving a request from a requesting entity to determine topics associated with a collection of documents, each document including a plurality of words and being represented as a set of (document, word) pairs, the document of the pairs being (i) and the words of the pairs being (j);

determining the topics associated with the collection of documents based on a hierarchy created by successively splitting nodes including a plurality of clusters, wherein each cluster reflects a topic and a document in the collection assignable to a set of clusters in the hierarchy based on different words included in the document, and wherein each cluster in the set is able to be associated with different paths in the



hierarchy;

storing a representation of the hierarchy in a memory; and

making the representation available to the requesting entity,

wherein in the foregoing steps, the (j) objects, for a given  $\alpha$ , are not collected in a product, rather a probability is determined such that the product is taken only after mixing over all classes  $\alpha$ , thus, different j objects are generated from different vertical paths of an induced hierarchy, that is, the paths in the hierarchy associated with non-null values of  $I_{i\alpha}$ , and all the instances of the hidden variable  $I_{i\alpha}$  obtain real values after a re-estimation using a modified EM process, and wherein as  $\alpha$  is able to be any node in the hierarchy, the (i) objects are able to be assigned to different levels of the hierarchy, accordingly, implementation of the model results in a pure soft hierarchical clustering of both (i) and (j) objects by eliminating any hard assignments of these objects.

15. (Currently amended) A computer program product, comprising a computer usable medium having a computer readable program code embodied therein, the computer readable program code adapted to be executed by a computer to implement a method

~~A computer-implemented method~~ for clustering a plurality of multi-word documents, wherein each document includes a plurality of words and is represented as a set of (document,word) pairs, represented as documents and words (i,j) into a hierarchical data structure including a root node successively split to produce ~~associated with~~ a plurality of sub-nodes, wherein each sub-node is associated with a topic cluster based on the plurality of documents, the method comprising:

retrieving a first document; associating the first document with a first topic cluster

based on a first portion of the first document;

associating the first document with a second topic cluster based on a second portion of the document; and

providing a representation of topics associated with the plurality of multi-word documents based on the hierarchical data structure including the first and second topic clusters,

wherein the first and second topic clusters are associated with a different sub-node,

wherein a probability of observing any pair of co-occurring objects represented as the set of (document,word) pairs, the document of the pairs being (i) and the words being (j), is modeled by defining a variable  $l_{i\alpha}$  which controls assignment of documents to the hierarchy, such that it is dependent on a particular (document,word) pair (i,j) under consideration during a topical clustering process, and a class  $\alpha$  ranges over all nodes in an induced hierarchy in order to assign a document (i object) to any node in the hierarchy, not just leaf nodes, and a class  $v$  is defined as an ancestor of  $\alpha$  in the hierarchy, the constraint on  $v$  ensuring that the nodes are hierarchically organized.

16. (Currently amended) The computer program product The method of claim 15, wherein the first and second portions contain at least one unique word.

17. (Currently amended) The computer program product The method of claim 15, wherein associating the first document with a first topic cluster comprises:

assigning the plurality of multi-word documents to a first class;

setting a probability parameter to an initial value; and  
determining, for the first document at the value of the parameter, a probability of an assignment of the first document to the first topic cluster based on a word included in the first document and the first class.

18. (Currently amended) The computer program product The method of claim 15, wherein associating the first document with a second topic cluster comprises:

assigning the plurality of multi-word documents to a first class;  
setting a probability parameter to an initial value; and  
determining a probability of an assignment of the first document to the second topic cluster based on a word included in the first document and the first class.

19. (Currently amended) The computer program product The method of claim 15, wherein providing a representation comprises:

providing the representation after each document in the plurality of multi-word documents has been associated with to at least one topic cluster corresponding to a sub-node in the hierarchy, wherein any of the plurality of multi-word documents may be associated to more than one topic cluster based on different portions of the respective document.

20. (Currently amended) A computer-implemented method for clustering data reflecting users, represented as a set of (data, user) pairs, into a hierarchical data structure including a root node successively split to produce

~~associated with~~ a plurality of sub-nodes, wherein each sub-node represents an action that is performed on a document collection, comprising:

accessing a user data collection reflecting a plurality of users who each perform at least one action on the document collection, wherein each action may be unique;

performing a clustering process that creates the hierarchical data structure, wherein the clustering processing comprises:

retrieving a first user data, associated with a first user, from the user data collection,

associating the first user data with a first sub-node based on a first action performed by the first user on the document collection, and

associating the first user data with a second sub-node provided the first user data is based on a second action, wherein the first and second sub-nodes are associated with different descendent paths of the hierarchical data structure at the same time;

storing a representation of the hierarchical data structure in a memory; and

making the representation available to an entity in response to a request associated with the user data collection.

21. (Original) The method of claim 20, wherein each action in the one or more actions includes:

writing to, printing, and browsing the document collection.

22. (Currently amended) A computer-implemented method for clustering a plurality of

images based on text associated with the images, where each image is represented as a set of pairs (image, image feature) and (image, text feature), into a hierarchical data structure including a root node successively split to produce ~~associated with~~ a plurality of sub-nodes, wherein each sub-node represents a different topic, the method comprising:

- accessing an image collection;

- performing a clustering process that creates the hierarchical data structure,

wherein the clustering processing comprises:

- associating a first image with a first sub-node based on a first portion of text associated with the first image, and

- associating the first image with a second sub-node based on a second portion of text associated with the first image, wherein the first and second sub-nodes are associated with different descendant paths of the hierarchical data structure at the same time ;

- storing a representation of the hierarchical data structure in a memory; and

- making the representation available to an entity in response to a request associated with the image collection.

23. (Currently amended) A computer-implemented method for clustering customer purchases, represented as a set of (customer, purchase) pairs, into a hierarchical data structure including a root node successively split to produce ~~associated with~~ a plurality of sub-nodes, wherein each sub-node represents a group of customers who purchased the same type of product from one or more business entities, the method comprising:

accessing information associated with a plurality of customers who purchased various types of products from a plurality of business entities;

performing a clustering process that creates the hierarchical data structure, wherein the clustering processing comprises:

associating a first customer with a first sub-node based on a first type of product purchased from a first business entity, and

associating the first customer with a second sub-node provided the first customer is based on a second type of product that the first customer purchased from a second business entity, wherein the first and second sub-nodes are associated with different descendant paths of the hierarchical data structure at the same time;

storing a representation of the hierarchical data structure in a memory; and

making the representation available in response to a request associated with the customer data collection.

24. (Currently amended) The method of claim 31, wherein the representation defines the probability of a document as the product of the probability of the (document, word) pairs it contains.

25. (Previously Presented) The method of claim 24, wherein the product is calculated after mixing the document-word pairs over the clusters.

26. (Previously Presented) The method of claim 25, wherein mixing the (document, word) pairs over the clusters comprises a probability model of the form:

$$P(x) = \sum_c P(c)P(x | c)$$

wherein  $c$  is the group of clusters involved in the calculation, and  $x$  is a (document, word) pair.